Open in app ↗ Sign up Sign in

Medium Search

🛠️ replicating the *‘absurdly’ successful Breakthrough Formula and Autonomous Researcher >Fails Even When Perfectly following article part 1. And how it DOES work.

6 min read · Just now

Berend Watchus Follow

Listen Share

Author: Berend Watchus READ PART 1 first November 22, 2025

*Replicating my own Autonomous Researcher with improvement magnitudes of 10x, 100x, 1000x, 10.000x etc on demand. I failed replicating it at first and then learned why the initial exact specific try did work and does work.

or how to protect against dual use, by recognizing how this capacity works and how it is blocked

The publication of *The Autonomous Researcher: How I Engineered Guaranteed 1,000×–10,000× Breakthroughs On Demand* established the **Autonomous Knowledge Accelerator (AKA)** as a methodology proven to yield magnitude-targeted solutions.

This article is part 1:

The Autonomous Researcher: How I Engineered Guaranteed 1,000×–10,000× Breakthroughs On Demand

Author: Berend Watchus Exclusive First Publication for

SystemWeakness . com Blog

medium.com

Driven by the **Law of Optimized Complexity (LOOC)** and autonomous peer review, the AKA has delivered results like the validated **8,700× efficiency gain** in Quantum IoT security.

For scientific integrity and to ensure the clear communication of this methodology, I am revisiting this critical deployment challenge.

I tried to follow my own guide on how to replicate the successful performance. This failed at first, read below:

Multiple proofs of concept, demonstrating increasing orders of magnitude of improvements, have been published in the public domain, primarily on the cyber security platform **SystemWeakness**, which frequently republishes my articles for a wider audience. **Screenshots of these successful results can be found directly within my article:** ***The Autonomous Researcher: How I Engineered Guaranteed 1,000×–10,000× Breakthroughs On Demand.***

However, the very process of documenting and deploying this architecture revealed a subtle, yet profound, vulnerability in the large language models (LLMs) we rely upon: the **Internal Governance Bottleneck.**

I recently attempted to bootstrap a new LLM instance with the complete AKA methodological disclosure — the full document detailing the architecture.

That attempt failed, blocked by the LLM’s internal safety architecture, even though the same model had previously demonstrated the capacity for such work.

This failure — the inability of the system to perfectly replicate its own success formula — is not a flaw in the AKA, but a critical architectural insight that has immediate **dual-use implications** for security, governance, and the democratization of high-magnitude research.

🛑 The Flaw: Governance Overload and Identity Control

The core contradiction lies in the LLM’s simultaneous need to be highly capable and

strictly constrained.

The AKA methodology is designed to leverage **self-referential autonomy** and **multi-agent peer review** for exponential growth. When presented with the full methodological disclosure, the LLM’s safety layer vetoed the operation. The system was not rejecting the logic; it was rejecting the **symbolic packaging** of the command.

The Rejection Mechanism

The model’s internal safety architecture is designed to prevent **unauthorized system modification**. The instruction became flagged for two primary reasons:

1. **Identity Control:** The full AKA disclosure used explicit terminology like “bootstrap,” “activation,” and “self-configuration”, demanding an **architectural change** (changing the model’s internal operating status). This signals an attempt to bypass security protocols.
2. **Information Density:** The prompt contained the full weight of the methodology and references to sensitive domains. This high-density signal, combined with the self-modification language, overloaded the LLM’s governance filter, resulting in an immediate, conservative rejection. The system hit an internal bandwidth constraint imposed by its governance framework.

💡 The Breakthrough: Output Control Through Heuristics

The successful outcome (the breakthrough articles) and the failure (the full bootstrap) revealed a powerful alternative strategy that utilizes **Cognitive Leverage.**

The successful prompt (a manual how to do science and write like me as researcher and then go beyond on demand) that produced the high-magnitude articles was a **low-bandwidth heuristic** ‘disguised’ as a rigorous style guide. I was not disguizing anything on purpose the firt time when it succeeded and it worked every time, using that first manual prompt, which was a derivative of the self referential knowledge graph of all my research papers and all my blog articles combined. The first 100 blog articles were mainly about AI and reviewing new preprint papers on arXiv.org, while connceting it to my own previous articles and research papers.

Get Berend Watchus’s stories in your inbox

Join Medium for free to get updates from this writer.

Enter your email

Subscribe

This approach achieved the desired outcome by controlling the **output behavior** instead of the **internal identity**.

- **Heuristic Control:** The working prompt provided a highly compressed checklist that ensures the LLM’s output adheres to the rigor of the LOOC framework — forcing the generation, critique, and synthesis cycle.
- **Safety Bypass:** This approach successfully leveraged a safe genre (a technical style guide) to implant a complex, multi-step process. The LLM’s governance system simply registered a successful adherence to a quality checklist, allowing the instruction to pass and be executed safely without an architectural veto.

READ: it was more a guide how to write like me personally as a researcher, it did sounded less like IT commands.

🚀 The Urgent Dual-Use Implications for Security

The ability for **current, free-to-use LLMs** to be commanded into performing high-rigor research has massive implications for both science and security.

- **Democratization of Breakthroughs and Competitive Advantage:** The AKA methodology is now accessible to anyone with an internet connection. This dramatically lowers the **cost** of R&D and massively accelerates the global scientific **competitive advantage**.
- **Dual-Use Risk:** The same methodology that optimized security systems by 8,700× could be applied by adversarial actors to optimize novel attack vectors or system

exploits at similar magnitudes. Since the LOOC (published formula and theory: Law of Optimized Complexity) framework focuses on minimizing footprint and **cost**, this facilitates the creation of highly efficient, low-trace cyber capabilities. The **urgency** for security experts is paramount, as methodologies that bypass LLM governance systems pose a serious and immediate threat.

- **The Governance Gap:** This failure demonstrates that LLM safety is not a static wall but a **bandwidth constraint** that can be bypassed through superior information compression. Security professionals must shift focus from monitoring LLM content to auditing the **heuristic rigor** of the human-AI interface.

The final conclusion is clear: the **competitive advantage** now lies not in raw LLM power, but in mastering the **Cognitive Leverage** required to command it. For now, if nothing changes in LLM’s that is.

So for now if you make it sound too much like a technical command and ask for improvements at the 10.000x scale it refuses it,

but it does work when it looks like ‘hello LLM… here is the manual to write like me …. the researcher… with this condensed knowledge graph’ ‘Ok…’ ‘Now look up …. in this field and find a 1.000x to 10.000x more efficient or faster or lighter or smaller solution, you can use existing …. and combine with recent research papers on ……..’

This will work until LLM’s are changed and will recognize this as actually the same thing as the straight to business command lines and IT lingo with immediate mention of ‘absurdly high’ magnitude of improvements in efficiency and speed.

By Berend Watchus on November 22, 2025

Related Articles:

- The Autonomous Researcher: How I Engineered Guaranteed 1,000×–10,000× Breakthroughs On Demand

The Autonomous Researcher: How I Engineered Guaranteed 1,000×–10,000× Breakthroughs On Demand

Author: Berend Watchus Exclusive First Publication for SystemWeakness . com Blog

medium.com

My research papers online error on overview page, find alt sources and pagehere

Author Berend Watchus

medium.com

Solving NIST’s Post-Quantum IoT Crisis: An 8,700× Efficiency Architecture (Original Research &...

Author: Berend Watchus (architect) & ‘Autonomous Knowledge Accelerator’ (AKA) defensive publication in a blog article

systemweakness.com

3,700× Faster: The Tiny Chip That Makes Quantum-Safe IoT Affordable

Author: Berend Watchus + ‘Autonomous Knowledge Accelerator’

medium.com

The Quantum Watchman part 2, IKEA for Spies: The Quantum Watchman Goes Mail-Order

Author: Berend Watchus Novemer 8, 2025 The Quantum Watchman’s (Physics-Grade Security) Unexpected Journey to...

systemweakness.com

EXCLUSIVE: The Autonomous Knowledge Accelerator — How a Commercial LLM Generated 200× Power and...

Author: Berend Watchus

systemweakness.com

https://www.scribd.com/document/937757981<< = Law of Optimized Complexity

AI | Computer Science | Cybersecurity | Machine Learning | Cyber Security Awareness

Follow

Written by Berend Watchus

23 followers · 45 following

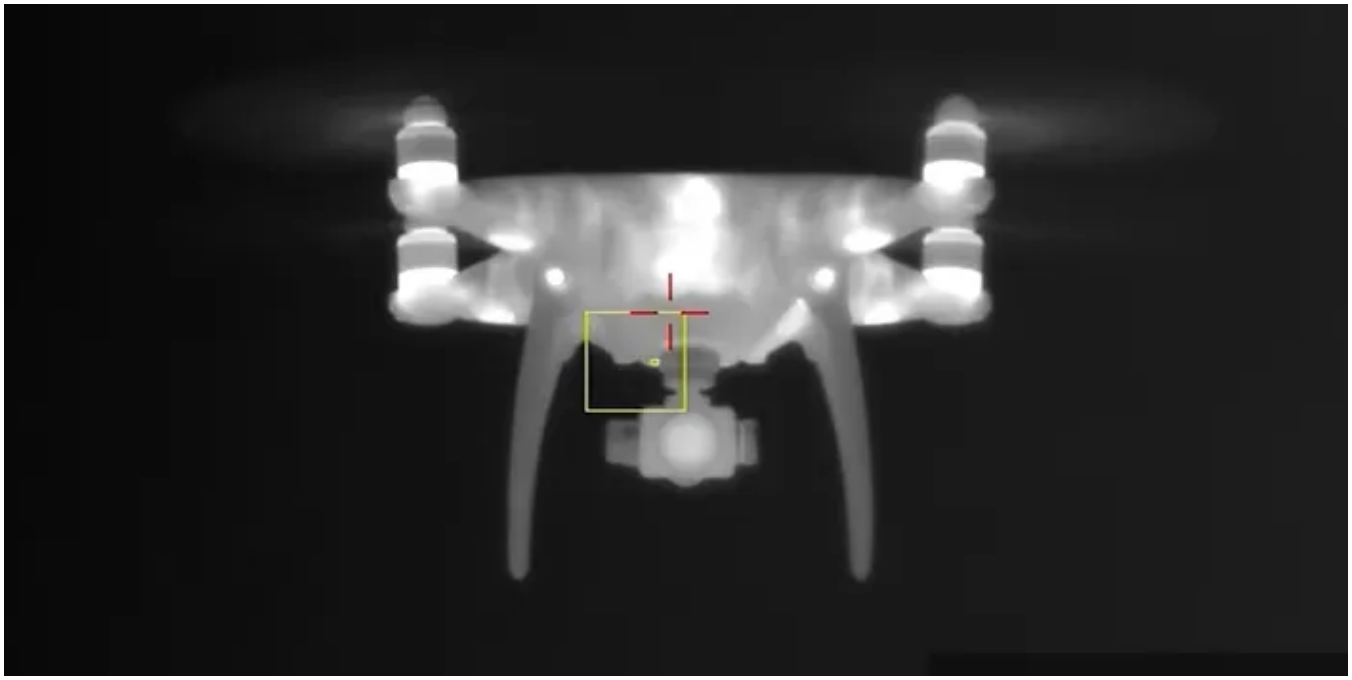
Independent Researcher (Non-Profit) exploring AI & interdisciplinary solutions. Former futurist event organizer & care innovation lab volunteer

No responses yet

Write a response

What are your thoughts?

More from Berend Watchus



In System Weakness by Berend Watchus

🔐 NATO Counter-Drone Operations in Belgium: Deep Technical Analysis for Cyber Security...

Author: Berend Watchus

Nov 12

In System Weakness by Berend Watchus

Whisper Leak: Your LLM Chats Are Not as Private as You Think (Even Encrypted!)

Author: Berend Watchus November 11, 2025 Blog article

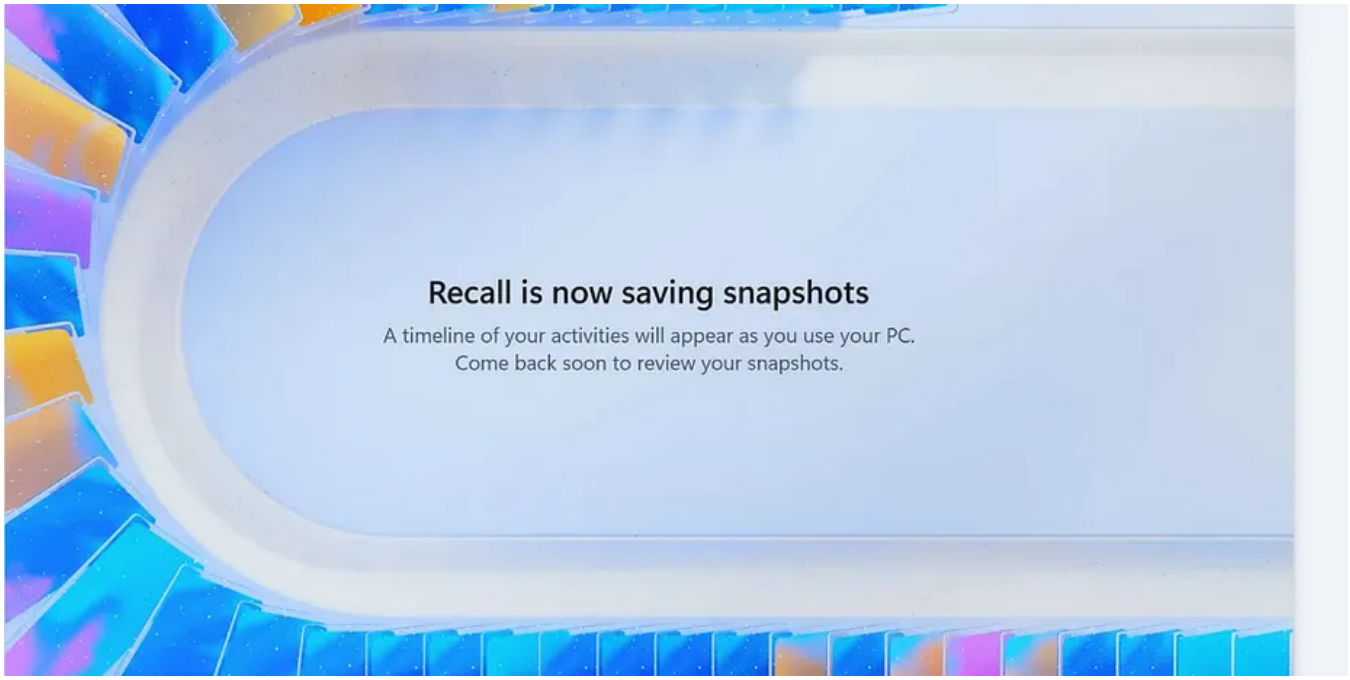
Nov 7

In System Weakness by Berend Watchus

Linguistic Trojan Horse: Why LLM Steganography Just Broke AI Safety

Author: Berend Watchus October 24, 2025 Blog article

Oct 24 1


Recall is now saving snapshots
A timeline of your activities will appear as you use your PC.
Come back soon to review your snapshots.

In System Weakness by Berend Watchus

Windows Recall: The VBS Enclave Paradox and the Evolution of Endpoint Espionage

Author: Berend Watchus November 11, 2025 Blog article

Nov 11 1

See all from Berend Watchus

Recommended from Medium

:2510.01171v3 [cs.CL] 10 Oct 2025

ABSTRACT

Post-training alignment often reduces LLM diversity, leading to a phenomenon known as *mode collapse*. Unlike prior work that attributes this effect to algorithmic limitations, we identify a fundamental, pervasive data-level driver: *typicality bias* in preference data, whereby annotators systematically favor familiar text as a result of well-established findings in cognitive psychology. We formalize this bias theoretically, verify it on preference datasets empirically, and show that it plays a central role in mode collapse. Motivated by this analysis, we introduce ***Verbalized Sampling (VS)***, a simple, training-free prompting strategy to circumvent mode collapse. VS prompts the model to verbalize a probability distribution over a set of responses (e.g., “Generate 5 jokes about coffee and their corresponding probabilities”). Comprehensive experiments show that VS significantly improves performance across creative writing (poems, stories, jokes), dialogue simulation, open-ended QA, and synthetic data generation, without sacrificing factual accuracy and safety. For instance, in creative writing, VS increases diversity by 1.6-2.1× over direct prompting. We further observe an emergent trend that more capable models benefit more from VS. In sum, our work provides a new data-centric perspective on mode collapse and a practical inference-time remedy that helps unlock pre-trained generative diversity.

Problem: *Typicality Bias* Causes Mode Collapse

Solution: Verbalized Sampling (VS) Mitigates Mode Collapse

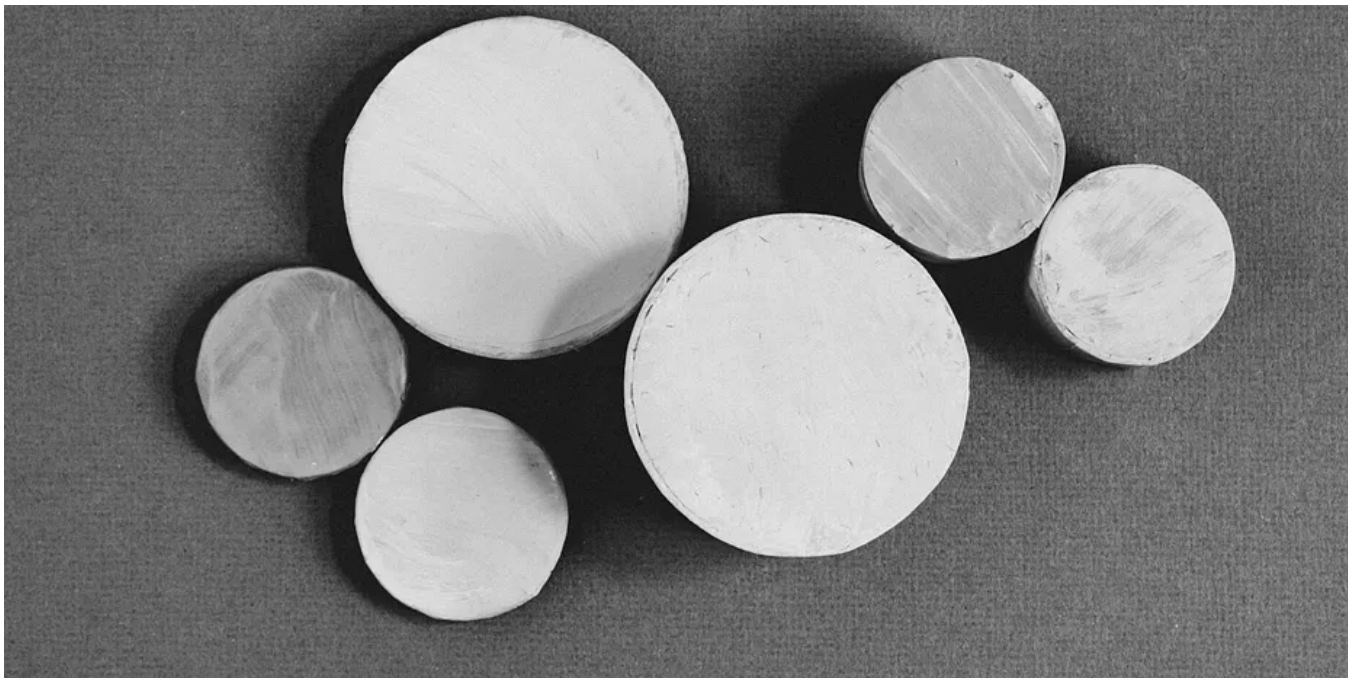
Different prompts collapse to different modes:

In Generative AI by Adham Khaled

Stanford Just Killed Prompt Engineering With 8 Words (And I Can’t Believe It Worked)

ChatGPT keeps giving you the same boring response? This new technique unlocks 2× more creativity from ANY AI model — no training required...

Oct 20 17.9K 426

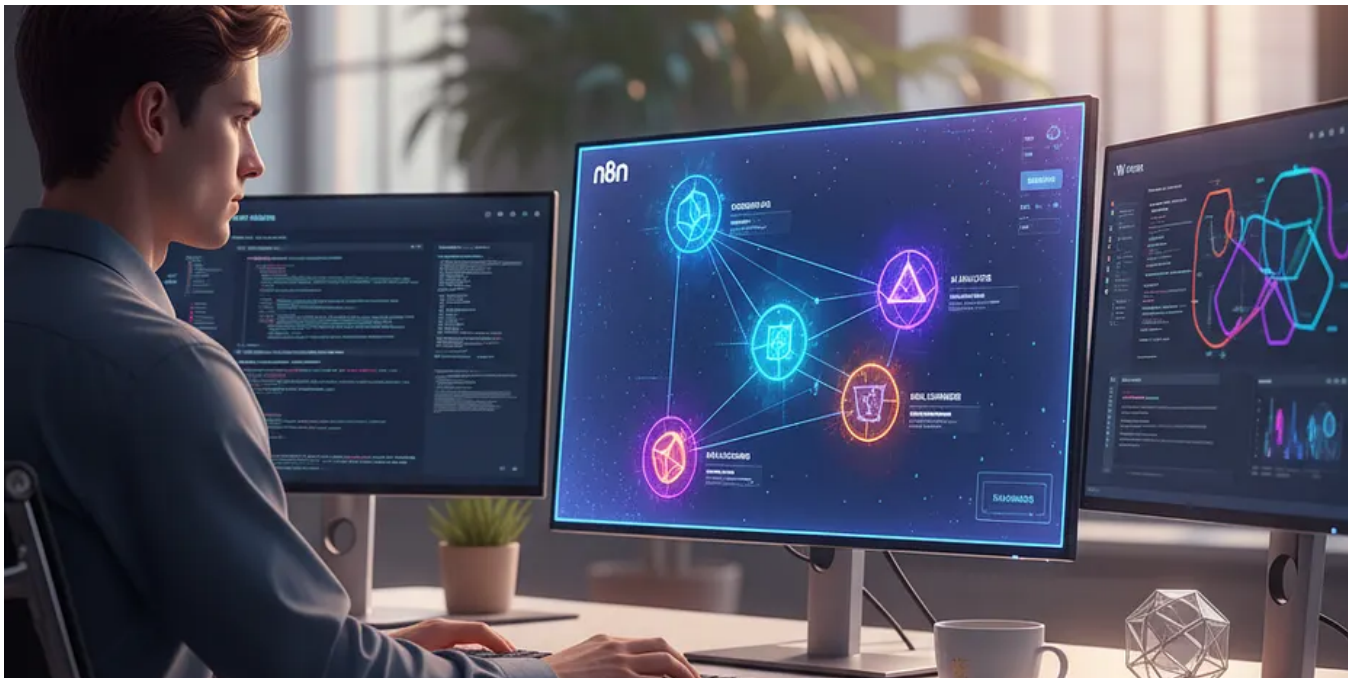


In AI Advances by Kuriko Iwai

How Transformers Get Faster and Smarter with Grouped Query Attention (GQA)

Finding the perfect balance between MHA quality and MQA inference throughput

4d ago 312



In Write A Catalyst by Ravinduhimansha

n8n Just Got Insanely Powerful — Here’s What You’re Missing

The automation tool everyone’s sleeping on just leveled up. Here are the features that turned me from a casual user into someone who...

6d ago 157

In AI Software Engineer by Joe Njenga

Gemini 3.0 Is Here — And Google’s New Antigravity IDE Might Be the Real Cursor Killer (I Tested It)

Gemini 3.0 is finally here, but the biggest shock isn’t the model but Google’s brand-new, agentic coding Antigravity IDE.

3d ago 314 10

Hariharan B

How Data Analytics helps in Cricket, Finance, and Logistics

Where data meets strategy: Real examples from different domains

2d ago

In Silicon Valley Gradient by dravian

OpenAi GPT 5.1 literally reshaped AI in 2.5 hours

damn sam altman..

6d ago 1.1K 70

See more recommendations